DOCUMENT RESUME

ED 086 517 SE 017 117

AUTHOR	Suydam, Marilyn N.
TITLE	Evaluation in the Mathematics Classroom: From What and Why to How and Where.
INSTITUTION	ERIC Information Analysis Center for Science, Mathematics, and Environmental Education, Columbus, Ohio.
PUB DATE	Jan 74
NOTE	70p.; Mathematics Education Reports
AVAILABLE FROM	Ohio State University, Center for Science and Mathematics Education, 244 Arps Hall, Columbus, Ohio 43210
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Annotated Bibliographies; Attitudes; Cognitive Development; Elementary School Mathematics; *Evaluation; *Evaluation Methods; *Instruction; *Mathematics Education; Secondary School Mathematics; *Test Construction; Testing

ABSTRACT

This document discusses the role and the scope of evaluation in the mathematics classroom. The scope of mathematics objectives to be evaluated, the scope of evaluation purposes in the mathematics classroom, and the scope of evaluation procedures are noted. Specific comments are made on various procedures: observations, interviews, inventories and checklists, attitude scales, and various types of paper-and-pencil tests. Both general and specific suggestions for planning tests and for writing various types of test items are included. An annotated list of selected references is included to direct attention to documents which will provide additional help. (JP)

FILMED FROM BEST AVAILABLE COPY

ED 026517

SMEAC

AN ERIC CENTER

SCIENCE, MATH, AND ENVIRONMENTAL EDUCATION INFORMATION REPORT

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

SMEAC/SCIENCE, MATHEMATICS, AND ENVIRONMENTAL EDUCATION INFORMATION ANALYSIS CENTER

. . . an information center to organize and disseminate information and materials on science, mathematics, and environmental education to teachers, administrators, supervisors, researchers, and the public. A joint project of The Ohio State University and the Educational Resources Information Center of USOE.

SE 017 117

ED 086517

MATHEMATICS EDUCATION REPORTS

Evaluation in the Mathematics Classroom: From What and Why to How and Where

by Marilyn N. Suydam

ERIC Information Analysis Center for
Science, Mathematics and Environmental Education
400 Lincoln Tower
The Ohio State University
Columbus, Ohio 43210

January 1974

Mathematics Education Reports

Mathematics Education Reports are being developed to disseminate information concerning mathematics education documents analyzed at the ERIC Information Analysis Center for Science, Mathematics, and Environmental Education. These reports fall into three broad categories. Research reviews summarize and analyze recent research in specific areas of mathematics education. Resource guides identify and analyze materials and references for use by mathematics teachers at all levels. Special bibliographies announce the availability of documents and review the literature in selected interest areas of mathematics education. Reports in each of these categories may also be targeted for specific sub-populations of the mathematics education community. Priorities for the development of future Mathematics Education Reports are established by the advisory board of the Center, in cooperation with the National Council of Teachers of Mathematics, the Special Interest Group for Research in Mathematics Education, the Conference Board of the Mathematical Sciences, and other professional groups in mathematics education. Individual comments on past Reports and suggestions for future Reports are always welcomed by the editor.

A major portion of the classroom teacher's duties involves the evaluation of student learning and skills. All too often, a discussion of evaluation procedures and techniques is clouded by a maze of tedious (if not bewildering) statistics and calculations. Statistics can be invaluable in analyzing and interpreting test results. But statistics are only as good as the scope, objectives, design and items of the test that generates the measures or numbers that enter into the calculations. Equating evaluation with statistical calculations is overlooking the most crucial and important aspects of evaluation.

This paper is designed to discuss these aspects of evaluation in a direct and simple manner. Addressed to the classroom teacher, it provides useful guidelines and techniques that will remove much of the mystery from the area of evaluation. An extensive annotated list of references provides sources of tests, item banks, and research on evaluation techniques as well as general references on evaluation in mathematics.

Jon L. Higgins
Editor

This publication was prepared pursuant to a contract with the National Institute of Education, U. S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, necessarily represent official National Institute of Education position or policy.

Table of Contents

	Page
Foreword	v
I. Introduction	1
II. The scope of evaluation	5
A. The scope of mathematics objectives to be evaluated	5
B. The scope of evaluation purposes	8
C. The scope of evaluation procedures	10
1. Observations	10
2. Interviews	11
3. Inventories and checklists	13
4. Attitude scales	14
5. Criterion-referenced tests	17
6. Norm-referenced tests	17
7. Standardized tests	18
8. Diagnostic tests	20
III. Developing tests	21
A. Planning the test	21
B. Writing the test items: some general suggestions	23
C. Short-answer questions or completion items	25
D. Multiple-choice items	26
E. True-false items	29
F. Matching items	30
G. Essay items	31

Page

H. Some related points 34

1. Item pools 34

2. Item analysis 34

3. Two definitions 37

IV. Concluding comment 38

Annotated list of selected references 40

Foreword

Many books have been written about evaluation and how to evaluate.

Some of these books go into a great amount of detail in developing the theses and the explanations for evaluation procedures and item construction.

Some of these books merely use a great many words.

This document neither goes into a great amount of detail nor uses a great many words. It is intended more as a quick reference guide than as an encyclopedia on the topic of evaluation of mathematics instruction. It should help the reader to review, to supplement, to develop questions. The list of references will help the reader delve further into the answers to these questions.

There are two emphases in this document:

(1) Evaluation is much more than paper-and-pencil tests: let's be aware of what each form of evaluation can contribute.

(2) Let's make each measure as good as possible.

The focus of what follows is on evaluation in the mathematics classroom: those evaluation procedures which are planned and administered and used in planning further instruction by the classroom teacher. Teachers also evaluate textbooks and materials and programs -- but we'll confine our attention to those instances where teachers and learners are interacting directly.

Evaluation in the Mathematics Classroom:
From What and Why to How and Where

I. Introduction

Imagine a classroom. Perhaps it's your classroom.

Imagine 25 or 30 students in that classroom. Perhaps they're your students.

Imagine the students sitting at desks.

*

Imagine you see the students clear everything off the tops of the desks, except for a pencil.

What did the teacher say at the point of the asterisk?

Imagine the sound of the teacher's voice. Insert the words the teacher is saying in place of the asterisk. The words are: "Clear your desks. Take out a pencil. You are now going to have a test."

When we think of evaluation in the mathematics classroom, tests come to mind immediately . . . tests where students sit at desks and write or circle or draw lines.

But is that all there is?

Imagine that same classroom three days ago. Groups of students are scattered around the room. Two are spinning a three-colored cube, and making a record of what color lands upward each time. Several are making a graph on a bulletin board. Others are stretching yarn

against various objects in the room. Some are seated with diagrams and worksheets, with games, with other materials before them.

Where is the teacher? What is he doing?

Is any evaluation occurring in the classroom at that moment?

Imagine the classroom four days ago. The students sit at their desks. The teacher is standing near the chalkboard. She writes some numerals on the board. She asks a question. Several students in turn respond. She asks another question. One student comes to the board and draws a diagram. The teacher queries the group by raising her eyebrows. Three students shake their heads "no", four nod "yes", the others look puzzled. The teacher asks another question.

Is any evaluation occurring in the classroom during this lesson?

Imagine the classroom five days ago. The students have moved their desks so they have tables grouped by fours. Each group follows the directions of a leader as they manipulate materials on their desks. They help each other; they talk about what they find happening. Then each records a response on a worksheet.

Is any evaluation occurring as they work on this lesson?

The answer to each question is obvious. If the teacher is teaching, the teacher is evaluating almost every minute on each of the days imagined -- and on any other day you want to imagine. Sometimes the evaluation leads to an immediate reaction: you smile

approval, or you frown; you say "good answer!", you say "that's on the right track"; you word a question so the student might see an error in his last response, you skip several questions because students are ready to move more quickly; you introduce a subtraction sentence instead of working only with objects, you get out rods as an alternative way of clarifying a mathematical idea. Sometimes the evaluation leads to notes on students' anecdotal records, a comment on a problem to pursue further, a change of lesson plans for next week.

Evaluation in the mathematics classroom consists of much more than a testing program involving paper-and-pencil tests on mathematical content. Measurement of the content goals of mathematics is comparatively easy: you can readily obtain an objective measure of certain computational skills and specific mathematical processes that form a portion of the mathematics curriculum. Measuring other goals of the mathematics curriculum is more difficult. Evaluation includes a wide variety of means of collecting evidence on students' behavior -- rating scales, questionnaires, checklists, reports from parents, student activities, and samples of students' work all provide useful evidence of behavior and progress. Observing, listening, presenting a task, interviewing: each makes a valid and viable contribution to the evaluation process.

But sometimes you evaluate with paper-and-pencil tests. Paper-and-pencil instruments have their place: they supplement other forms of evaluation. The very process of preparing for and taking a test helps the student to synthesize what he has learned. The responses to

specific items help the teacher to diagnose a weakness or confirm what he has observed in the day-by-day process of observing student reactions and behaviors. Both students and teachers take stock: this mathematical idea or fact or skill or concept has been mastered and can be used in developing newer content. Another mathematical idea or fact or skill or concept needs to be given more thought or practice or development.

One of the purposes of this document is to help you to develop better paper-and-pencil measures. Tests are going to be a part of the educational environment for a long time to come, if only because they provide a feasible way of finding out, in a relatively short amount of time, what or how well each child is learning certain content. Tests yield concrete and detailed evidence economically and in convenient form. Tests are, however, only tools whose value lies not in mere use but in the skill and understanding of the user. Good tests do not just happen: they require much thought and careful planning.

Another purpose of this document is to review other possible approaches to evaluation. And to provide a guide to some of the pertinent literature on evaluation in mathematics education, an annotated listing of selected references is provided. Occasionally, numbers in parentheses have been inserted in the text to direct you to a reference; other references are included on the list without being cited.

II. The scope of evaluation

Evaluation is a continuing, integral aspect of mathematics teaching, concerned with the improvement of instruction. Evaluation ascertains whether the teacher is teaching what he thinks he is teaching and the learner is learning what the teacher thinks the learner is learning. Evaluation is qualitative as well as quantitative: it involves appraisal as well as measurement, for it includes the stage of making value judgments.

Evaluation takes a variety of forms, since there is no one technique that is equally appropriate for measuring all aspects of learning. Both cognitive factors and affective factors must be assessed: the feeling and the doing aspects as well as the knowing and thinking aspects.

A. The scope of mathematics objectives to be evaluated

Scope-and-sequence charts in textbooks and curriculum guides provide one way of determining the dimensions of the mathematics program. Some mathematics educators have described the scope in various ways; for example:

> *In the study of mathematics a student must learn facts, develop concepts, use symbols, and master processes and procedures. But he should also learn to develop generalizations and to sense the presence of mathematical ideas and structures not only in abstract situations but also in many areas of human activity. He should develop his reasoning powers in order to prove or disprove a statement by deduction or to predict an event with appropriate probability. It is the function of evaluation to determine how well a student has mastered these varied aspects of mathematics.*
>
> (Sueltz, 14, pp. 7-8)

Other writers have developed models to aid in the process of designing instructional materials and tests. The taxonomy developed

by a committee working with Bloom provides a basic paradigm for the analysis of educational goals in general (3, 4, 16); other models have been developed which are specific to the goals of mathematics education. Bloom's Taxonomy is presented in terms of two domains, the cognitive and the affective; the cognitive domain, not surprisingly, has been of most concern to those evaluating mathematics, even though the importance of the affective domain is recognized. Goals in the cognitive domain have been categorized into six main categories, plus many subcategories:

1. Knowledge -- recognition or recall of specific material
2. Comprehension -- grasping the meaning of material
3. Application -- using information in concrete situations
4. Analysis -- breaking down material into its parts
5. Synthesis -- putting together parts to form a whole
6. Evaluation -- judging the value of material and methods for given purposes

In the affective domain, there are five categories: receiving, responding, valuing, organization, and characterization by a value.

In one adaptation of Bloom's taxonomy for mathematics education, five categories are considered; evaluation is incorporated as a component of the analysis and synthesis categories (1). A taxonomy was developed specifically by the School Mathematics Study Group for use in evaluating mathematics achievement in the National Longitudinal Study of Mathematical Abilities (2, 30, 76). It details four levels of the cognitive domain: computation, comprehension, application, and analysis.

These models have each been of inestimable aid to curriculum

developers and test constructers. Yet many teachers find it difficult to recall the levels, and even more difficult to apply them. Pikaart and Travers (72) attempted to simplify the model so that it would really help teachers to describe specific learning goals, yet be comprehensive, flexible, and functional. They provide for three dimensions -- goals or products, content, and teacher behavior or processes, including planning, teaching, and evaluation. For the goal dimension, they consider both cognitive and affective facets; they note that in practice it is difficult to distinguish activities that are planned for one or the other: cognitive and affective goals are interrelated and interwoven in instruction. Thus the same model may be considered for both facets:

1. Knowledge
    a. Statements
    b. Basic skills
2. Understanding
    a. Concepts
    b. Principles
3. Problem Solving
    a. Formulating hypotheses and testing them
    b. Proving theorems
    c. Solving non-routine problems

Levels are important to consider in setting goals and developing objectives for instruction, in planning instructional activities and procedures, and in evaluating instructional outcomes. Too frequently mathematics evaluation encompasses only the lowest level -- knowledge. It is easy to construct an objective test at the knowledge level; it is much more difficult to construct tests and other evaluation

procedures that assess higher cognitive levels. A model can aid in teaching, even if only by making everyone aware of the need to evaluate higher-level outcomes.

B. The scope of evaluation purposes

Each teacher evaluates for at least three purposes:

1. To assess the mathematics program in the classroom and in the school.

The success of your mathematics program is not determined by how well it compares with the program in other schools. The important concern is the impact it has on helping your particular students to learn mathematics. Is the content appropriate for your students? How well are they progressing toward the mathematical goals you have set? Are they able to apply their knowledge and skill in new situations? Does the program make the students want to continue to learn more mathematics? Do they enjoy doing and using mathematics? Is the content important and worthwhile mathematics? Is the program teachable and learnable?

Comparisons with other students in other schools can help you to attain some perspective on how well your students are doing. The National Assessment of Educational Progress and various state assessment programs are another attempt to provide such perspective (55, 67, 68, 90, 114, 118, 127, 136, 147, 149). But you are not teaching "other students in other schools". Your goal must be to help each of the students in your classroom to learn and to enjoy mathematics as well as he is able.

2. To assess the achievement of the students in each classroom.

We have discussed this in general in the introduction to this document, and we'll discuss it more specifically in the sections to follow this one. The vital factor to note is that you must evaluate students in terms of both progress and status. Testing supplements other evaluation procedures as a means of ascertaining how well students have succeeded in mastering important content and acquiring important skills.

3. To diagnose individual strengths and weaknesses.

You can use test results to place students in instructional materials, to group students for instruction, to assign grades. You can also use them to help you to learn more about how to teach more effectively.

Far too many mathematics tests consist simply of examples for which students are to provide answers. Far too often these tests are corrected by a check for correct and incorrect answers. The teacher who merely obtains the total score made by a student on a test is overlooking the greatest value of the test for instructional purposes. Alas -- so much is thrown in the wastebasket! Analysis of how the student reached the correct or incorrect answer can tell you far more than mere knowledge of whether the answer was right or wrong. Analysis of how individual questions were answered can tell you more than a total score can.

Evaluation procedures other than tests are invaluable in providing diagnostic information. As you listen and observe, you build the basis for interpreting test scores and deciding how to structure

your teaching.

C. The scope of evaluation procedures

This section contains comments on various types of evaluation techniques: first, non-paper-and-pencil procedures, then paper-and-pencil instruments.

1. Observations

Many mathematics lessons have a component in which students work in small groups or individually on tasks, assignments, or worksheets. This is a time when evaluating students' mathematical behavior is of singular importance. You can move about the room, observing students as they work, listening as they talk among themselves, making notes, questioning, making suggestions. You also observe during discussion periods, but your involvement in the discussion sometimes keeps you from attaining perspective: then you need to use your evaluation immediately as you continue the discussion. You have little chance to make notes. Your primary purpose is to guide. When you are free to observe as children work independently, you can evaluate even more effectively, with a defined perspective, and you can limit your observation to specific aspects of student behavior.

Note the method of attacking problems used by a student, and how he proceeds to work a problem. Note the expression on his face, his mannerisms, his concentration. Note how consistently he works, where he meets difficulty, when he becomes careless. Observe the emotional climate of the room. Observe the student's level of independence. Does he really need your help when he raises his hand,

or does he need encouragement or praise? How dependent is he on help from you, from textbooks, from other students? Does he try various ways of solving a problem, or does he try to apply the last procedure used in class?

Make a simple memo that describes the situation and the behavior you've observed -- an anecdotal record. Use a small notepad or cards.

Name	*Date*	*Situation*	*Behavior*	*Comment*
Sue	1|17	group lesson, developing meaning of fractions with graph paper	quick to help neighboring students	
	1|20	computation game	missed most combinations in which she had to multiply by 7 or 9	redevelop and practice multiplication with 7 and 9

File the anecdotal records in the student's folder, in which you also place examples of his daily work, project reports, and other papers (79).

Sometimes audiotape (or videotape) can be used to provide a record that you can go back over and analyze in more detail than when you are involved with the group. Photographs can provide a record of project work and "products". You can compare progress with more objectivity than simply through memory of what was done.

2. Interviews

An interview is an attempt to remove the restriction of writing, both that involved in your development of a test item and that of the child in developing an answer. You can delve more precisely

into how a student solves an example or problem. You can learn how he goes about finding an answer. You can follow as he describes what he is doing and why (56, 65, 69, 81).

Basically, the interview procedure is simple:

(1) Face the student with a problem.

(2) Let him find a solution, as he tells you what he's doing.

(3) Challenge him, to elicit his highest level of understanding.

Present the student with an example written on a card:

$46\overline{)327}$

Have him explain the procedure he follows while computing the answer.

Make notes as he works: sometimes it's helpful to have an exact record of what he says. Challenge him with such questions as, "Are you sure that's correct?" "What if I said the answer was ___?" "Is there any other way you could find the answer?" And remember that the two most important questions in an interview are "How?" and "Why?".

Other suggestions for interviewing include:

(1) Establish rapport and maintain a relaxed atmosphere. The student needs to understand what he is to do. You don't want him to search for the answer he thinks you want -- you want his answers, not yours. And you want to know what he's thinking. You want him to respond naturally, freely, and fully.

(2) Select your examples and questions for your purpose. At times, you'll interview only some students; at other times, the whole class. Use more than one example of a particular type, to determine

how consistently he works.

(3) Don't teach: don't give answers, and avoid leading questions and suggestions. Do as little talking as you can. You want to find out what the student is thinking.

(4) Record the student's answers and thinking and whatever he does, as you go. You may want to write fast, or tape record, or categorize or code, using an interview. Don't rely on memory to make a "true" record after the interview is over. Careful records will enable you to ascertain patterns and provide other evidence for diagnostic teaching.

(5) Time may be a problem, or it may be an excuse. The mathematics laboratory or open classroom can facilitate interviewing -- time is more flexible, students are more "available". But if the teacher is serious about using interviewing as a means of finding out more about what students have learned and are learning, the time can be found -- when others have a worksheet, during free-reading time, etc. Schedule time one day a week, or some time each day.

(6) You may want to have a student use a tape recorder without you being present. Have him tell how he does some aspect of mathematics, why he attacks a problem as he does, why he likes or dislikes mathematics. A group of students might discuss various ways of solving a problem. You can play the tape back later and analyze student thinking more carefully and from a different perspective than you can if you're involved in the interview.

3. Inventories and checklists

An inventory is a check of what the student knows about a specific

topic or what he knows about the total program. It's probably especially useful at the beginning of the year. In oral form, primary-level teachers find it an indispensible alternative to a written test. At upper levels, it may be written and administered just as any other test is. The inventory frequently is used to survey the previous year's work or the status of students (both individuals and class) as they begin work in your classroom. Such a test is an aid in assessing the readiness of students for more advanced work, as well as a diagnostic aid. List the items and skills you want to inventory. Decide how you will inventory each: what direction will you give the student, what tasks and materials will you use, or what test items will you need.

A checklist is a type of inventory: a list of kinds of behavior to look for -- for example, evidence of interest in mathematics, applying mathematics, working with others, using a range of materials, etc. Rating scales are like checklists but provide for a degree of appraisal: *turns in assignments: never -- occasionally -- always*

or *counts on fingers: frequently -- sometimes -- never*

4. Attitude scales

We believe that the affective component of learning is important: if we are interested in and enjoy mathematics, we'll learn it better. Attitudes involve both cognitive and non-cognitive aspects, an intellectual appreciation and emotional reactions. Thus attitudes toward mathematics involve many facets, ranging from awareness of the structural beauty of mathematics and of the important roles of mathematics to feelings about the difficulty and challenge of learning

mathematics to interest in particular type of mathematics or particular methods of being taught mathematics.

We attempt to assess the student's attitudes toward mathematics in several ways. One primary way is through observation: by observing his expressions, comments, and behaviors as a student reacts in a mathematical situation, we infer how he feels about mathematics. We note how often he chooses a mathematical activity when he has an option, how readily he attempts to apply mathematical ideas to real-life situations, how enthusiastically he reacts in a mathematics lesson. We can use a checklist as a systematic approach to recording observations.

At times we ask the student to comment directly on his attitudes. We have him write an essay on a question such as, "Do you generally like or dislike mathematics? Why or why not?" Or we have him complete sentences such as "I like mathematics because --- ". We may ask him to rank in order of preference the subjects which he is studying: we infer the level of his preference for mathematics by where he places it in relation to other subject areas.

Perhaps the most widely used measure of attitudes is the attitude scale (91, 101, 125). Half a dozen scales have been extensively used; on many of them, items such as those on the scale on the next page appear. The scale attempts to ascertain, less directly and therefore hopefully with greater reliability or credibility, how strongly the student likes or dislikes mathematics.

You can construct your own scale to measure specific aspects of mathematics; the procedure is concisely outlined by Corcoran and Gibb (in reference 14).

Attitudes Toward Mathematics

(Scale Form B)

Marilyn N. Suydam and Cecil R. Trueblood
The Pennsylvania State University

This is to find out how you feel about mathematics. You are to read each statement carefully and decide how you feel about it. Then indicate your feeling on the answer sheet by marking:

A - if you strongly agree
B - if you agree
C - if your feeling is neutral
D - if you disagree
E - if you strongly disagree

1. Mathematics often makes me feel angry.
2. I usually feel happy when doing mathematics problems.
3. I think my mind works well when doing mathematics problems.
4. When I can't figure out a problem, I feel as though I am lost in a mass of words and numbers and can't find my way out.
5. I avoid mathematics because I am not very good with numbers.
6. Mathematics is an interesting subject.
7. My mind goes blank and I am unable to think clearly when working mathematics problems.
8. I feel sure of myself when doing mathematics.
9. I sometimes feel like running away from my mathematics problems.
10. When I hear the word mathematics, I have a feeling of dislike.
11. I am afraid of mathematics.
12. Mathematics is fun.
13. I like anything with numbers in it.
14. Mathematics problems often scare me.
15. I usually feel calm when doing mathematics problems.
16. I feel good toward mathematics.
17. Mathematics tests always seem difficult.
18. I think about mathematics problems outside of class and like to work them out.
19. Trying to work mathematics problems makes me nervous.
20. I have always liked mathematics.
21. I would rather do anything else than do mathematics.
22. Mathematics is easy for me.

(Attitudes Toward Mathematics scale, continued)

23. I dread mathematics.
24. I feel especially capable when doing mathematics problems.
25. Mathematics class makes me look for ways of using mathematics to solve problems.
26. Time drags in a mathematics lesson.

5. Criterion-referenced tests

Paper-and-pencil instruments can help as you evaluate the individual student in terms of his own progress: what has he learned that he didn't know before you taught that unit on fractions or the metric system or binomials? You compare the performance of a student with his previous performance. You design a test to ascertain whether or not each student has learned what you have taught. You set a level that says, if he gets this percentage of the items correct, adequate mastery of the topic can be assumed. You can also ascertain how well your class has mastered a particular topic, so the test parallels the work in class. These are criterion-referenced tests or mastery tests (98, 99, 115, 129, 130, 142).

6. Norm-referenced tests

Paper-and-pencil instruments can also provide you with information on the status of the student in relation to other students in the class. A student is compared with others, his achievement evaluated relative to the achievement of the class. The test may also be designed in terms of ascertaining whether students have been learning what you think they should be learning from your teaching. But instead of setting a mastery level, a scale is used: you expect a few students to do very well, a few to do poorly, but most to attain

an "average" level. These tests are based on the content you have taught as are criterion-referenced tests, but they're norm-referenced measures.

7. Standardized tests

Another form of norm-referenced test is used in almost every classroom at least once a year: the commercially-published standardized test (52, 83). (A few standardized tests are criterion-referenced, but most are norm-referenced. Occasionally teacher-developed tests are standardized by large school districts, by developing standards for their own students.) Standardizing a test refers to developing prescribed, uniform requirements for administration and scoring, and to the statistical analysis after the test is given to a large, specified group of students, resulting in the development of norms. With the use of norms based on what students in many classrooms have scored, you have a measure of how well your students are learning when compared with many others.

Standardized tests are not a substitute for teacher-made tests but a complement. More careful preparation and research are provided than it is ordinarily possible for any individual teacher to provide for his own classroom tests. The content has been determined on the basis of common elements of widely used courses of study and textbooks. Care must be taken to ascertain that the standardized test adequately covers the expected outcomes of your school's mathematics program. Aspects that are unique to your program will not be included, and you'll have to make provision for testing them. Many producers of standardized tests publish outlines of test content to compare with your local program.

Some guidelines have been suggested for selecting a standardized test:

(1) Formulate clearly the purposes that will be achieved by use of the test: precisely what kinds of information are the tests expected to supply? What outcomes are to be measured? What use is to be made of test results?

(2) What tests are available that will meet your needs? Lists of tests are available (23, 24, 25, 26) and should be consulted.

(3) Obtain copies of those tests which, from their descriptions, appear to meet your purposes. (Most test publishers will furnish sample test materials.)

(4) Examine the tests and the test manuals for appropriateness for your particular needs, reliability, ease of administration and scoring, kinds of normative data provided, and evidences of careful development. Norms should have been established in schools similar to yours. There should be at least several thousand students in the norm group if the norms are to be accepted with confidence. The norm should be stated in a convenient form, such as percentiles (which indicate the percentage of students whose performance is found to be below any score) or grade norms (which show how well the average student in a specified grade has performed). The manual should include explicit directions for administration and suggestions for interpreting and using the results. Make sure that the time requirements are reasonable in terms of your school.

It seems safe to state that no students can avoid standardized tests as they progress through school. Therefore it is wise to teach students how to take such tests: just reading the standardized

test directions as they begin the first test is not enough. Develop tests that use the same types of items that will be met on standardized tests. This is particularly necessary for young children: many rarely see a multiple-choice item, for instance, until it is met on a standardized test.

8. Diagnostic tests

Some standardized tests are planned to be specifically diagnostic (19, 44). They usually cover a limited scope in much greater detail than a test of general achievement. They are arranged to give scores on the separate parts.

You can also develop a teacher-made test that is diagnostic. The value of this type of test will depend on its ability to reveal specific weaknesses in the achievement of individual pupils. When you have identified the point at which the student begins to have difficulty, you can begin to help him to overcome the difficulty. Knowing that the student attained a score of thirty per cent on a division test provides you with little guidance on how to improve your instruction; knowing that the student attained an incorrect answer to 673 ÷ 4 tells you little more. But knowing that the student's answer to that example was 16 remainder 3 tells you that perhaps he needs help in understanding the placement of the answer in the quotient, that perhaps he needs help with place value, that perhaps he does not understand the algorithm. It provides you with some information to follow up on.

In developing a diagnostic test, select the examples with care: they must be examples which readily allow errors of the types you predict. Have the student show all of his work -- even when you use

multiple-choice or other types of items.

III. Developing tests

In this section some suggestions for developing tests will be considered. These suggestions have been drawn from many sources (e.g., 26). An attempt has been made to be comprehensive, but you must look elsewhere for elaboration and illustrations. Some general procedures will be given first: these apply to the planning and development of all types of instruments. Then some specific suggestions to consider in developing various types of items will be presented.

A. Planning the test

A well-planned test must be designed to accomplish the purpose it is to serve. Have the kinds of information that you hope to get from the test clearly in mind.

1. List the objectives to be assessed by the test.

Consider: what have you taught? What mathematical content and ideas are really important for the students to have learned? Test objectives should correspond to instructional objectives; instructional objectives suggest the type of evaluation procedure and test item to use. Remember that some objectives are best measured by non-paper-and-pencil procedures.

The objectives will vary in scope and number depending on the type of test. For a mastery test, it may be that each objective toward which you taught will be assessed by several questions. For an achievement test at the end of a longer period of time, you must be more selective in choosing only the major critical points, those

which are important in the hierarchy or as a "base" for future learning.

2. Decide on the types of items to be constructed.

The type of item depends on the nature of the objective to be measured. Once you have determined that an objective can be measured adequately by a paper-and-pencil item, you need to decide what type of item to use. Some mathematical objectives are measured well by short-answer or completion items, or by multiple-choice items; a few objectives are best measured by true-false or matching items. Such objective-type items (so-called because they can be scored objectively, with independent scorers obtaining the same results) measure knowledge and comprehension levels efficiently. A relatively large field of content can be sampled, for objective-type items can be answered quickly and one test can contain many questions. This broad coverage helps provide a reliable instrument. For higher-level outcomes, consider essay tests.

3. Decide on the number of items to be written for each objective.

There are no simple rules for determining the "right" number of items to use for measuring each objective. The content of a test should reflect the relative amount of emphasis each objective has received in the actual instruction: thus the number of items will be in proportion to the amount of emphasis. The level of the items will be similarly related to the objectives. Take into consideration whether the interpretation of results will be in terms of each separate objective or the test as a whole. And of course consider the amount of time available for administration of the test.

To help ensure that the completed test will give each objective the desired coverage, develop an outline of specifications to serve as a guide for item construction.

content (objectives)	*% of emphasis in instruction*	*number of items*	*level of items* K	U	*upper*
forming equivalent classes	10	4	1	2	1
adding 'like' fractions	20	8	2	3	3

Tests should measure an adequate sample of the learning outcomes and content included in the instruction. You can never ask all of the questions you might like to: you can only test a sampling of the most important outcomes.

B. Writing the test items: some general suggestions

The role of each item is to ascertain whether a student has attained the objective or not. There should be nothing about the structure or presentation of an item that leads those who know the correct answer to get the item wrong or those who do not know the answer to get the item right.

1. Select the measurement technique that is most effective for the specific objective.
2. Use clear, simple statements. Use language that students understand. Choose concise vocabulary, and sentence construction that is appropriate to the level of your students. Break a complex sentence into two or more separate sentences.
3. Design each item so that it provides evidence that an objective has been achieved. Avoid testing for unimportant details, unrelated bits of information, or irrelevant material.

4. Check items against the table of specifications to make sure that you have the desired emphasis on various content objectives at various levels of difficulty.

5. Work with another teacher or group of teachers in reviewing each others' items. Cut out points of doubtful importance or correct unclear wording.

6. Adopt the level of difficulty of a test item to the group and to the purpose for which it is to be used.

7. Initially, you may want to write more items than you will need on the final form of the test. Then you can discard weaker items. Many teachers write down items each day for possible inclusion on a test, to help ensure that important points will not be omitted.

8. Have each student work from a separate copy of the test, rather than from a test written on the chalkboard.

9. Number all items consecutively from the first item on the test to the last.

10. Avoid putting part of an item on the bottom of one page and the rest on the top of the next page.

11. If the form of a test or a group of items is unfamiliar, use sample items to help clarify the directions. Spend some time teaching students how to take a test.

12. Precede each group of items with a simple, clear statement telling how and where the student is to indicate his answers.

13. When you want the student to show his computation, provide adequate space near each item. "Boxing in" this space helps you to locate it quickly.

14. Begin a test with easy items. Placing difficult items at the beginning of a test is likely to discourage average and below-average achievers. You can then arrange items so that the test gets increasingly more difficult, or you can mix easy and difficult items.

15. Many times you'll need to have more than one type of item on a test (short-answer, multiple-choice, etc.). Place all items of one kind together. Always have more than one or two items of a particular type (except possibly of the essay type).

16. Avoid a regular sequence in the pattern of responses: students are likely to answer correctly without considering the content of the item at all.

17. Eliminate irrelevant clues and unnecessary or non-functional clues, but provide a reasonable basis for responding.

18. Make directions to the student clear, concise, and complete. Instructions should be so clear that each student knows what he is expected to do, although he may be unable to do it.

19. Prepare a key containing all the answers that are to be given credit. Make it so that it can be placed beside the answer spaces used by the students.

20. After the test, go over questions with your students: they can point out ambiguities and other errors, helping you to improve items for future use.

21. Analyze student responses to each item, for diagnostic use.

C. Short-answer questions or completion items

The short-answer item employs a question, an incomplete statement, or a computational example to elicit from the student appropriate words, symbols, or numbers. It is generally limited to questions that call for facts: who, what, when, where, how many. Many classroom mathematics tests are solely of this type: it is frequently used to measure the ability to compute. You can present a number of computational exercises, or you can focus the student's attention on particular aspects of a computation.

In the completion item, certain important words or phrases are replaced by blanks to be filled in by the student. It must be very carefully prepared, or it is likely to measure only rote memory, or intelligence rather than achievement.

1. State the item so that only a single brief answer is required and possible.

2. Use a direct question when possible; switch to an incomplete statement only when greater conciseness is possible.

3. Words to be supplied should relate to the main point of the statement.

4. Blanks should be placed at the end of the completion statement.

5. Avoid giving extraneous clues to the answer.

6. If the answer can appear in more than one form, give specific directions about which form to use. Indicate such things as the degree of precision for numerical answers and whether labels must be used.

7. Avoid the use of sentences taken directly from the textbook. They are frequently ambiguous out of context, and encourage rote memorization.

8. Do not give clues to the answer by varying the number or length of the blanks.

D. Multiple-choice items

The multiple-choice item consists of a stem which is a question or an incomplete sentence presenting a problem situation, followed by several alternatives, which are possible solutions to the problem. One of the alternatives is the correct answer; the others are plausible answers, called distracters because their function is to distract students who are uncertain of the correct answer. The stem may also be a problem, graph, or diagram followed by the alternatives relating to it.

The ease of scoring undoubtedly plays a big part in the popularity of multiple-choice items. Student answers are easy to read and unambiguous. The use of computer-scoring has made the multiple-choice item virtually the only type used when a computer is available or for standardized tests. In general, scores on multiple-choice tests are comparable to those that would be obtained from free-response tests, for the same level of content.

But there are other reasons for deciding to use multiple-choice items: they tend to provide a more adequate measure of many objectives than do other objective-type items. Multiple-choice tests have high

reliability compared with other types of tests. And with careful analysis and development, the multiple-choice item can be adapted to most types of content and to most levels of objectives. It can assess the student's ability to recognize facts or relationships, to discriminate, to interpret, to analyze, to make inferences, to solve problems. Its biggest weakness is that it allows the student to guess, but this affects scores less than on other types of items.

Multiple-choice items should not be used when a simple question is adequate, that is, where there is clearly only one correct answer and no plausible distracters. They should not be used when there are only two plausible responses; a true-false item is usually effective in that instance.

1. Make directions explicit, so that the student knows exactly what type of response is required. Is more than one answer possible? Is he to select "the correct answer" or "the best answer"? How should he record his answer? Should he guess if he isn't sure of the correct answer?

2. The stem should present a single worthwhile problem to be solved, expressed clearly and without ambiguity. State the question so there can be only one interpretation. Check on the clarity of the stem by covering the alternatives and determining whether the question could be answered without the choices.

3. Make each question independent of other questions. Students are often able to select the correct answer to one item because of information gleaned from another item. Where an answer to one item is used in succeeding items, students who miss that item will miss the succeeding items.

4. Make alternative choices as brief as possible. Instead of repeating words in each alternative, include them in the stem.

5. State the stem in positive form whenever possible. When negative wording is used, emphasize it by underlining or by capitalizing.

6. The best alternative choices to the correct answer are those using commonly mistaken ideas or common misconceptions or errors commonly made by the students. Excellent distracters can be obtained from

incorrect responses on short-answer, completion, or essay tests.

7. In general, use the same number of alternatives for each item on a test. But remember that an item is not improved by adding an obviously wrong answer merely to obtain another alternative. Generally four or five alternatives are used, to reduce the chance of guessing the correct answer. It is better to have only four alternatives when five plausible choices are not available.

8. Make all incorrect responses equally plausible or "attractive" to the student who does not know the correct answer. If plausible distracters are difficult to find, use another type of item rather than ineffective alternatives. The more homogeneous the alternatives, the more difficult the item will be. The correct answer is one which cannot be refuted.

9. Make all alternatives grammatically consistent with the stem, and parallel in form. Avoid verbal clues which might enable students to select the correct answer or to eliminate an incorrect alternative: similarity of wording in the stem and the correct answer, for instance, or including two responses that are all- inclusive or two that have the same meaning. Check the structure by reading each alternative with the stem.

10. Do not consistently make the correct response longer or shorter than the distracters. Unconsciously, there is a tendency to include the greatest amount of detail in the correct answer.

11. Avoid the use of qualifying words such as "always", "never", or "all" as much as possible: they are clues to a test-wise student that an alternative probably is not true.

12. Avoid use of the alternative "all of the above" and use "none of the above" with care. The inclusion of "all of the above" makes it possible to answer the item on the basis of partial information: the student can realize that it is the correct choice by noting that two of the alternatives are correct, or that it is not the correct choice by noting that at least one of the alternatives is incorrect. His chance of guessing the correct answer is thus increased. The use of "none of the above" may be measuring only the ability to detect incorrect answers: he may do this and still not know the correct answer. If you want to reduce the chances of students estimating the answer without doing an entire computation (when that is the objective), use a completion-type item.

13. Avoid using a pattern for the position of the correct response. Students are quick to perceive patterns or apparent patterns and select their answers accordingly. Use some system of random order for the positions of the correct answers on each multiple-choice test -- and check to make sure that patterns did not inadvertently occur. Many teachers fail to use a, d, and e, as

often as they use b and c as distracters. Students learn that their chances of guessing the correct answer are better if they guess b or c. Be sure the correct response is placed in all positions approximately the same number (but not exactly the same number) of times.

14. Control the difficulty of the item either by varying the problem in the stem or by changing the alternatives.

15. Use an efficient item format.
    a. List alternatives on separate lines, one under the other, making them easy to read and compare.
    b. Use letters in front of alternatives, to avoid confusion with numerical answers.

E. True-false items

The true-false item can be difficult to construct, for statements must be unquestionably true or false. To construct such items to measure important outcomes is difficult: they adapt best to the measurement of specific facts, understanding of principles or generalizations, and common misconceptions. They can be used only when there are only two possible alternatives. Because they are highly subject to guessing, true-false items have little value as diagnostic tools.

"Alternative-response items" are variations in which the student must respond "agree" or "disagree"; "right", "partly right", or "wrong"; or with similar words. Other variations include items in which attention is directed to an underlined word or phrase; after deciding that any statement is false, the true words are to be inserted in place of the underlined words. Students can also be asked to state why the statement is true or false. Cluster true-false items deal with a single idea; such mathematical content as graphing can be tested with such an item, where students are asked to look at a graph and then respond to a series of true-false items about the data portrayed.

1. Have students circle T and F, or write T and F or + and 0 (rather than t and f or + and -, which cannot be distinguished as readily).

2. State the item clearly and specifically so that it is unequivocably true or false. Avoid the use, however, of specific qualifiers such as "always" or "never" -- or use them in both true and false statements. Check for ambiguities.

3. The item should deal with a single definite idea. The use of several ideas in each statement tends to be confusing and the item is more likely to measure reading ability than achievement. There should be no more than one problem-setting clause.

4. Avoid making true statements longer than false statements.

5. Make the crucial element readily apparent to the student. It is better to have the crucial element come at the end rather than in the early part of a two-part statement.

6. Have an approximately equal (but not exactly equal) number of true and false statements (vary the proportions from test to test).

7. Randomly arrange true and false items; check to be sure there is no inadvertent pattern.

8. Avoid trick statements which appear to be true but are really false because of some inconspicuous or trivial word or phrase.

9. Avoid statements that are partly true and partly false.

10. Avoid the use of statements extracted from textbooks. Out of context, such statements are often unclear or ambiguous.

F. Matching items

The matching item measures ability to discriminate between several items of similar material as they are related in a given way with items of another set. The matching exercise is essentially a modification of the multiple-choice form. When all of the responses in a series of multiple-choice items are the same, the matching format is more appropriate. Said another way, unless all of the responses in a matching item serve as plausible alternatives for each premise, the matching format is inappropriate.

Matching items can be used for such content as definitions and words defined, measurement and formulas, or geometric shapes and names. They are most appropriate for testing at the knowledge level; it is difficult to adapt them to testing for comprehension and higher-level goals.

1. Place the premise column on the left, the briefer responses on the right. Each of the items in the left column should have a test item number; the responses should be preceded by letters. Have the student place his answer to each item in a space to the left of the item number.

2. The items in the two columns must be homogeneous (that is, no responses should be logically excludable as answers by a student who is uninformed). If they are not homogeneous, the student may be provided with clues which will help him to match the terms, resulting in easier test items. Selection of the correct match should be dependent on knowledge of the correct answer, not on ability to eliminate incorrect answers on the basis of extraneous information.

3. To reduce the effect of guessing, one column should contain more terms than the other. Directions should clearly indicate whether responses may be used once, more than once, or not at all.

4. Do not include too many items in either column: a maximum of twelve items in the premise column should be considered. Longer lists require too much searching time. There should be more responses than items in the premise column when responses are to be used only once, to avoid the selection of the last response on the basis of elimination.

5. Place the items in the response column in some logical order, to enable the student to scan the list quickly to find the term he has in mind. Jumbling the terms merely increases searching time.

6. Be sure that there is only one response which is the correct match for each premise when responses are to be used only once.

G. Essay items

Essay items are not often used on mathematics tests, but they can and should be. Such items require the student to do more than compute a solution or recall specific facts. He must think about

mathematics and its meaning. He must organize his own ideas and express himself effectively in his own words, using both knowledge and reasoning. Purely factual information is not assessed as efficiently as with objective-type items, but higher levels of reasoning can be tapped. Essay questions can be used to assess comprehension, application, and analysis outcomes; they provide a means of assessing a student's ability to synthesize or to evaluate mathematical ideas which is rarely provided by objective items. Essay questions that assess complex achievement are apt to include such key words as why, explain, compare, relate, interpret, criticize, develop, derive, classify, illustrate, and apply.

There are difficulties in using essay items, as you're aware. An essay test covers a limited field; the questions take so long to answer that relatively few can be answered in a given period of time. A representative sampling of content is not feasible. Essay items are subjective, more difficult to score, and less reliable than objective-type items. Scores are apt to be distorted by writing ability and by bluffing. The student who is fluent can often avoid discussing points of which he is unsure. But there are things you can do to minimize these problems, beginning with the writing of clearly defined items -- general enough to offer some leeway, but specific enough to set limits.

1. Use essay questions to evaluate achievement only on those objectives which are not readily tested by other types of items.

2. Phrase the questions as precisely as possible and be specific in wording, so the objective of the item is clear and the student is made aware of the specific scope or limits to be included in the answer.

3. Make clear to the student the basis on which the answer will be judged, such as content, organization, comprehensiveness, relevance, appropriateness, etc.

4. Require all students to answer all questions, so they are all taking the same test.

5. Indicate suggested time allotments for each question. Be sure that the student has time to write adequate answers: time must be allowed for thinking as well as for writing. Provide adequate space for answers.

6. Discuss ways of answering essay questions with the students.

Since scoring essay items can be difficult, here are some suggestions which will increase objectivity.

1. List specific objectives for each essay question as it is written. Evaluate in terms of the objectives. Separate scores may be given for style of writing or spelling, but should not contaminate the evaluation of the mathematical objective being assessed.

2. Write out the essentials of a complete answer to each question or prepare a model answer ahead of time. Use it in the same way in scoring each paper. This does not preclude adding other acceptable points made by students. Determine the number of points to be assigned to each part of the model answer, or determine criteria for levels of expected quality.

3. Keep the identity of students unknown. Have the student use a coded numeral on his paper or have him write his name on the back or at the end of the test.

4. Read one question through the entire set of papers, scoring each item for all papers before going on to the next item.

5. More uniform standards can be applied by reading the answers twice. At the first reading, sort the papers into several piles. Then reread to check on the uniformity of answers in each pile and make any necessary changes in rating. Assign the same item score to all papers in a pile.

6. Reshuffle the papers so that a paper may not be scored unduly high or low because of its position, after scoring each item.

H. Some related points

1. Item pools

An item pool is simply a collection of test items that you can put together in various combinations to form a test. Several items may be developed for testing each specific objective; you can select the one that best meets test requirements. You'll probably find that a card file is the easiest way of filing the items. Write each item on a card, noting the topic or objective in one corner. At the bottom or on the back, record what you've learned about the item: when it should be used, what percentage of students get it correct each time you use it, and so on.

Other sources of models for items include commercial tests, textbooks for students or teachers, collections of items or item banks (31, 32, 33, 37, 38, 39, 40, 41, 42, 62, 110, 111, 112, 131, 132, 133, 134), and the tests which were constructed for various research studies (145).

Item sampling, incidentally, is a technique for assessing the status of a group of students (45, 75, 86, 97, 102, 108, 116, 130, 138). Since your focus is usually on how well students are achieving, rather than on how well content is being achieved across students, you will probably not use item sampling techniques. You may find the term appearing more frequently in various articles about testing, however.

2. Item analysis

Item analysis is the process of studying the students' responses to each item. An item analysis can tell you how difficult an item is and how well each question discriminates between high- and low-

ranking students. It's especially important if you are going to re-use the item: it can indicate whether or not an item needs to be revised. It's also useful even if you don't plan to use the item again, for it can tell you what questions are especially appropriate to test certain objectives. Or it can be used simply as part of your diagnostic procedures.

Computer programs are used for item analysis for tests that are developed for research studies, for standardized tests, and for other tests that will be used by many groups of students. For most classroom tests, only simple item analysis procedures seem warranted. Here are several suggestions (29):

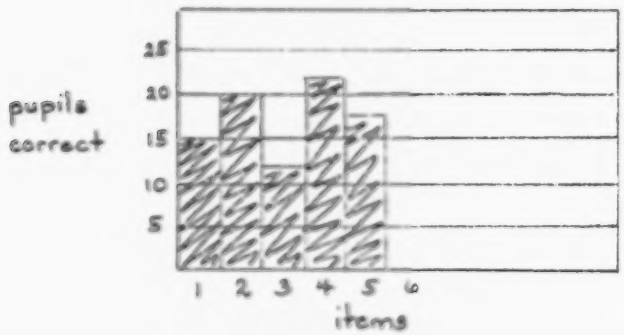
(1) Look at the test: what items were missed by many students? Were they missed because of a "fault" in the item or was there a "fault" in the instruction? What do you do next? Revise the item or revise the instruction.

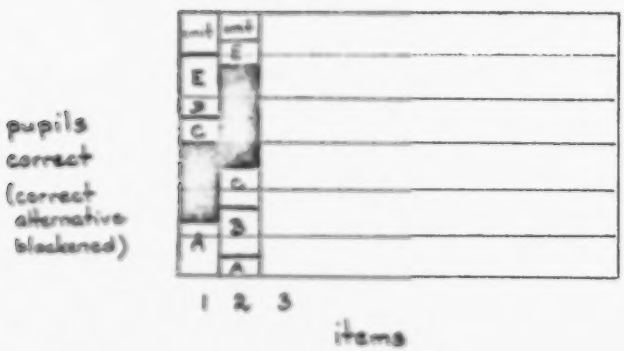
(2) A simple measure of difficulty is the percentage of students who got the item correct. This gives you an approximation of how difficult the item is. By recording this information for each item in your item pool, you can build a test which will be at an appropriate difficulty level. This is especially helpful when you're developing a test in which you want to rank students; each item should then be of medium difficulty -- approximately 40% to 60%. (For mastery tests, your standards will be different.)

You can check the students' papers yourself to obtain the percentages, or you can do an item analysis by a show of hands. Call out the item numbers one by one, and have each student who has the item correct hold up his hand. Count and record the number of hands.

Have students convert it to a per cent, or do it yourself.

You can extend this activity by building a graph with the students, recording either the number of students who got the item correct or the number of incorrect responses. (For multiple-choice items, keep a record of the number selecting each alternative.)





Are there any patterns in the graph? What items were missed most? Are there areas involving any particular objective?

(3) To do a more sophistocated item analysis, use this procedure:

(a) Arrange the test papers in order from highest to lowest score.

(b) Select the highest one-third and the lowest one-third (approximately), setting aside the middle one-third of the papers.

(3) For each item, count the number of students in the upper group who got it right and the number in the lower group who got it right. Let's say you have 10 papers in the upper one-third and 10 in the lower one-third. For one item, here's the count for the correct answer: *upper -- 7*

lower -- 3

(4) Convert these numbers to percentages:

$$\frac{7 + 3}{20} \text{ (total students)} = 50\%$$

If the item is a good one for ranking students, then substantially more students in the upper group will have answered correctly. The harder the item, the lower the percentage getting it correct. Items on which many more students in the lower group got the item correct need revision.

(5) On multiple-choice tests, determine the effectiveness of distracters, by comparing the number of students in upper and lower groups who selected each incorrect alternative. A good distracter will attract more students from the lower group than from the upper group. Each distracter should attract some students or it is not serving effectively as a distracter. (Different criteria, however, apply to mastery tests.)

3. Two definitions

Any test, whether constructed by an individual teacher or commercially published, should meet several criteria, including acceptable validity and reliability. Validity pertains to the relevance of the test. Are you collecting the right kinds of information? Does the test measure the skills, understanding, or knowledge that

it was intended to test? Does it measure the significant behaviors that are specified in the objectives? Are all items relevant to those behaviors? Is the test a balanced sampling of the behaviors? Reliability pertains to the consistency of the test. How accurate and stable is the test? Does it measure the same achievement consistently? The nature of mathematics helps to make mathematics tests quite reliable. If a test were perfectly reliable, the students would have the same score or be ranked in the same order if the test were repeated, or a parallel form of the same test were administered. Reliability is commonly reported by a coefficient of correlation between forms of the test or between two halves of the same test. Perfect reliability is represented by a coefficient of 1.00. Usually a coefficient of at least .80 is expected on an objective mathematics test; many mathematics tests have reliabilities of .90 and higher. Tests of computational ability are usually more reliable than tests of problem-solving ability.

You probably have many other questions. Answers to these questions, whether about definitions or about testing or about other aspects of evaluation, may be answered by one or more of the references included at the end of this document. Each reference is annotated to provide you with a clue to its contents.

IV. Concluding comment

The goal of evaluation is improving instruction. Measuring or assessing or testing only indicates: the teacher then has to do something as a result of what he's learned. This document has not

attempted to consider the most difficult task in teaching: the use of the knowledge and understanding gained from evaluation. Evaluation is only a beginning . . . you must continue the process of teaching.

Annotated Listing of Selected References

I. Books: evaluation in mathematics education

1. Avital, Shmuel M. and Shettleworth, Sara J. Objectives for Mathematics Learning: Some Ideas for the Teacher. Bulletin No. 3. Toronto: Ontario Institute for Studies in Education, 1968.

Developing and classifying objectives is discussed, with specific suggestions for the teacher.

2. Begle, Edward G. (Ed.). Mathematics Education. Sixty-ninth Yearbook, National Society for the Study of Education. Chicago: University of Chicago Press, 1970.

Chapter 9 (J. F. Weaver) is a discussion of evaluation in terms of the role of the classroom teacher. Chapter 10 (Edward G. Begle and James W. Wilson) presents the NLSMA design as a model for the evaluation of mathematics programs.

3. Bloom, Benjamin S. (Ed.). Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain. New York: David McKay, 1956.

This classic book presents a model for classifying and developing objectives in the cognitive domain, with many illustrative test items for varied subject areas.

4. Bloom, Benjamin S.; Hastings, J. T.; and Madaus, G. F. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill Book Co., 1971.

Evaluation problems which a teacher is likely to encounter, and a framework and techniques for test construction are discussed in detail. Illustrations of objectives, testing techniques, and sample test items for specific subject areas (including mathematics) are presented.

5. Buffie, Edward G.; Welch, Ronald C.; and Paige, Donald D. Mathematics: Strategies of Teaching. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

One chapter focuses on the evaluation of mathematics instruction.

6. Butler, Charles H. and Wren, F. Lynwood. The Teaching of Secondary Mathematics (4th ed.). New York: McGraw-Hill Book Co., 1965.

In one chapter, procedures for evaluation are considered.

7. Collier, Calhoun C. and Lerch, Harold H. Teaching Mathematics in the Modern Elementary School. New York: Macmillan Co., 1969.

Chapter 14 includes discussion of testing, interviews, observations, and checklists, with sample items.

8. D'Augustine, Charles H. Multiple Methods of Teaching Mathematics in the Elementary School (2nd ed.). New York: Harper & Row, 1973.

Chapter 17 includes a discussion of inventories and tests.

9. Davis, Frederick B. Educational Measurements and Their Interpretation. Belmont, California: Wadsworth Publishing Co., 1964.

This is a general textbook on measurement procedures.

10. Dutton, Wilbur H. Evaluating Pupils' Understandings of Arithmetic. Englewood Cliffs, New Jersey: Prentice-Hall, 1964.

Suggestions are given for evaluating mathematical understanding in various ways, through a planned program.

11. Ebel, Robert L. Measuring Educational Achievement. Englewood Cliffs, New Jersey: Prentice-Hall, 1965.

This general textbook on evaluation provides specific suggestions.

12. Grossnickle, Foster E.; Brueckner, Leo J.; and Reckzeh, John. Discovering Meanings in Elementary School Mathematics (5th ed.). New York: Holt, Rinehart and Winston, 1968.

In one chapter, methods of evaluation are discussed, with illustrations of what techniques to apply for particular objectives.

13. Johnson, Donovan A. and Rising, Gerald R. Guidelines for Teaching Mathematics (2nd ed.). Belmont, California: Wadsworth Publishing Co., 1972.

One chapter provides suggestions on various evaluation procedures.

14. Johnson, Donovan A. (Ed.). Evaluation in Mathematics. Twenty-sixth Yearbook, National Council of Teachers of Mathematics. Washington: The Council, 1961.

This yearbook contains discussions on all phases of evaluation, ranging from theoretical considerations to specific suggestions for developing items.

15. Kramer, Klaas. The Teaching of Elementary School Mathematics (2nd ed.). Boston: Allyn & Bacon, 1970.

Chapter 21 includes discussion of observation, tests of various types, with some sample items, and interviews.

16. Krathwohl, David R.; Bloom, Benjamin S.; and Masia, Bertram B. Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook II: Affective Domain. New York: David McKay Co., 1964.

A model for classifying and developing objectives in the affective domain is presented, with suggestions for curriculum and instruction.

17. Marks, John L.; Purdy, C. Richard; and Kinney, Lucien B. Teaching Elementary School Mathematics for Understanding (2nd ed.). New York: McGraw-Hill Book Co., 1965.

Chapter 14 includes a typical sample program for evaluation, followed by discussion of evaluation in mathematics programs.

18. Riedesel, C. Alan. Guiding Discovery in Elementary School Mathematics. New York: Appleton-Century-Crofts, 1967.

Chapter 15 discusses all phases of evaluation in mathematics, with specific suggestions and illustrative items.

19. Reisman, Fredricka K. A Guide to Diagnostic Teaching of Arithmetic. Columbus, Ohio: Charles E. Merrill Publishing Co., 1972.

Specific case studies are discussed, with some explicit suggestions and diagnostic tests.

20. Spitzer, Herbert F. Teaching Elementary School Mathematics (4th ed.). Boston: Houghton Mifflin Co., 1967.

Chapter 14 includes a discussion of observation, testing, and other evaluative procedures.

21. Sund, Robert B. and Picard, Anthony J. Behavioral Objectives and Evaluational Measures: Science and Mathematics. Columbus, Ohio: Charles E. Merrill Publishing Co., 1972.

Examples of various types of behavioral objectives and hints on developing instruments to measure progress toward these objectives are included.

22. Stanley, Julian C. Measurement in Today's Schools (4th ed.). Englewood Cliffs, New Jersey: Prentice-Hall, 1964.

This is a general text on evaluation, with specific suggestions for the development of testing procedures.

II. Books: tests and test construction for mathematics education

23. Braswell, James S. Mathematics Tests Available in the United States (3rd ed.). Washington: National Council of Teachers of Mathematics, 1972. (pamphlet)

This comprehensive but brief listing of tests includes title, author, grade levels and forms, availability of norms, publisher, reference, and publication dates.

24. Buros, Oscar Krisen (Ed.). The Seventh Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1970.

This invaluable two-volume reference lists all tests published (including 96 mathematics tests), with critical reviews of each test by specialists in the field. Six previous yearbooks provide information on earlier tests.

25. Buros, Oscar Krisen (Ed.). Tests in Print. Highland Park, New Jersey: Gryphon Press.

This is a comprehensive listing of tests.

26. Gronlund, Norman E. Constructing Achievement Tests. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1968.

Specific suggestions for developing tests are presented, with illustrative items based on the content of the book.

27. Horn, Dorothy M. The Writing of Multiple-Choice Mathematics Test Items. Toronto: Ontario Institute for Studies in Education, 1970. (pamphlet)

The steps in constructing test items are considered and suggestions for revising poorly constructed items are given.

28. Husén, Torsten (Ed.). International Study of Achievement in Mathematics, Volumes I and II. New York: John C. Wiley and Sons, 1967.

These volumes include reports on the intentions and background of the International Study, the hypotheses, the mechanics of test construction, test and attitude inventory scores and correlations, interpretations, and item statistics.

29. Katz, Martin (Ed.). Evaluation and Advisory Service Series: No. 3, Selecting an Achievement Test: Principles and Procedures; No. 4, Making the Classroom Test: A Guide for Teachers; No. 5, Short-cut Statistics for Teacher-Made Tests. Princeon, New Jersey: Educational Testing Service. (pamphlets)

    Procedures for teachers to use are briefly described.

30. Romberg, Thomas A. and Wilson, James W. The Development of Tests. NLSMA Report No. 7. Stanford, California: School Mathematics Study Group, 1969.

    The purpose, scope, and development of the tests which were used in the National Longitudinal Study of Mathematical Abilities are described.

31. Wilson, James W.; Cahen, Leonard S.; and Begle, Edward G. (Eds.). X-Population Test Batteries, Parts A and B. NLSMA Report No. 1. Stanford, California: School Mathematics Study Group, 1968.

32. Wilson, James W.; Cahen, Leonard S.; and Begle, Edward B. (Eds.). Y-Population Test Batteries, Parts A and B. NLSMA Report No. 2 Stanford, California: School Mathematics Study Group, 1968.

33. Wilson, James W.; Cahen, Leonard S.; and Begle, Edward B. (Eds.). Z-Population Test Batteries. NLSMA Report No. 3. Stanford, California: School Mathematics Study Group, 1968.

    Tests and scales used in the National Longitudinal Study of Mathematical Abilities are included in the above three volumes. X-Population includes grades 4-8; Y-Population, grades 7-11; Z-Population, grades 10-12.

34. Wilson, James W.; Cahen, Leonard S.; and Begle, Edward G. (Eds.). Description and Statistical Properties of X-Population Scales. NLSMA Report No. 4. Stanford, California: School Mathematics Study Group, 1968.

35. Wilson, James W.; Cahen, Leonard S.; and Begle, Edward B. (Eds.). Description and Statistical Properties of Y-Population Scales. NLSMA Report No. 5. Stanford, California: School Mathematics Study Group, 1968.

36. Wilson, James W.; Cahen, Leonard S.; and Begle, Edward B. (Eds.). Description and Statistical Properties of Z-Population Scales. NLSMA Report No. 6. Stanford, California: School Mathematics Study Group, 1968.

Item statistics for the tests and scales used by NLSMA are included in the above three volumes.

37. (OISE). Experiences with Sets and Numbers: Mathematics Evaluation Materials Package Project. Toronto: Ontario Institute for Studies in Education, 1972. (booklet)

    This is a set of objectives and companion test items for mathematics in grades 4 through 6.

38. (OISE). Grades 7 and 8 Mathematics Item Pool. Booklet I: A Study of the Set of Whole Numbers. Booklet II: A Study of the Set of Fractional Numbers. Booklet III: Geometry and Measurement, A Study of Integers, Presentation and Interpretation of Data. Booklet IV: Supplementary Classroom Problems. Teachers' Guide. Toronto: Ontario Institute for Studies in Education, 1969.

    Each booklet contains a pool of items on the specified mathematical topic.

39. (SMSG). Kindergarten Test Batteries, Description and Statistical Properties of Scales. ELMA Technical Report No. 1. Stanford, California: School Mathematics Study Group, 1971.

40. (SMSG). Grade 1 Test Batteries, Description and Statistical Properties of Scales. ELMA Technical Report No. 2. Stanford, California: School Mathematics Study Group, 1971.

41. (SMSG). Grade 2 Test Batteries, Description and Statistical Properties of Scales. ELMA Technical Report No. 3. Stanford, California: School Mathematics Study Group, 197_.

42. (SMSG). Grade 3 Test Batteries, Description and Statistical Properties of Scales. ELMA Technical Report No. 4. Stanford, California: School Mathematics Study Group, 197_.

III. Articles: evaluation and tests in mathematics education

43. Baker, Eva L. The Effects of Manipulated Item-Writing Constraints on the Homogeneity of Test Items. Journal of Educational Measurement 8: 305-309; Winter 1971.

    Subtraction items written to meet behavioral objectives and under certain specifications were easier than those written under no specifications or for non-behavioral objectives.

44. Burns, Paul C. Analytical Testing and Follow-up Exercises in Elementary School Mathematics. School Science and Mathematics 65: 34-38; January 1965.

Suggestions for diagnostic testing and teaching are given.

45. Cahen, Leonard S.; Romberg, Thomas A.; and Zwirner, Walter. The Estimation of Mean Achievement Scores for Schools by the Item-Sampling Technique. Educational and Psychological Measurement 30: 41-60; Spring 1970.

The item-sampling technique was found to be satisfactory, with the precision of estimation increasing as the number of students tested in a school increased.

46. Caldwell, Edward. Group Diagnosis and Standardized Achievement Tests. Arithmetic Teacher 12: 123-125; February 1965.

Simple item analysis procedures are noted, with suggestions for using test results in diagnosing problems.

47. Carry, L. Ray. A Critical Assessment of Published Tests for Elementary School Mathematics. Arithmetic Teacher 21: 14-18; January 1974.

Characteristics of norm- and criterion-referenced standardized tests are discussed, with specific comments on 19 tests.

48. Cliffe, Marian C. The Place of Evaluation in the Secondary School Program. Mathematics Teacher 49: 270-273; April 1956.

Why, what, and how to evaluate in secondary-school classes in mathematics are discussed.

49. Coppedge, Floyd L. and Hanna, Gerald S. Comparison of Teacher-Written and Empirically Derived Distractors to Multiple-Choice Test Questions. Journal for Research in Mathematics Education 2: 299-303; November 1971.

Teachers did not produce multiple-choice geometry item-distractors that were very similar to the discriminating errors students made using completion format. It was suggested that multiple-choice tests be developed after analysis of completion tests.

50. Elder, Florence L. Using "Take-Home" Tests. Mathematics Teacher 50: 526-528; November 1957.

The use of take-home tests is briefly discussed; suggested test items are included.

51. Epstein, Marion G. Testing in Mathematics: Why? What? How? Arithmetic Teacher 15: 311-319; April 1968.

Types of tests and how to plan and develop tests are discussed, with several illustrative items.

52. Epstein, Marion G. Standardized Tests Can Measure the Right Things. Mathematics Teacher 66: 294, 363-366; April 1973.

Arguments to support the use of standardized tests are presented. The other side of the issue is discussed by Wilson (1973).

53. Erlwanger, S. H. Benny's Conception of Rules and Answers in IPI Mathematics. Journal of Children's Mathematical Behavior 1: 7-26; Autumn 1973.

A problem which can occur in the multiple-test situation of an individualized program is described.

54. Fellows, Martha M. A Mathematics Attitudinal Device. Arithmetic Teacher 20: 222-223; March 1973.

Two forms of a scale to ascertain children's feelings about mathematics are given.

55. Foreman, Dale I. and Mehrens, William A. National Assessment in Mathematics. Mathematics Teacher 64: 193-199; March 1971. (Similar article in Arithmetic Teacher 18: 137-143; March 1971.)

Procedures for developing exercises for National Assessment are reviewed. Content categories are listed, but no items are included.

56. Gray, Roland F. An Approach to Evaluating Arithmetic Understandings. Arithmetic Teacher 13: 187-191; March 1966.

An individual interview inventory was developed to measure varying levels of understanding in multiplication.

57. Hammitt, Helen. Evaluating and Reteaching Slow Learners. Arithmetic Teacher 14: 40-41; January 1967.

The use of evaluation sheets with slow learners is suggested.

58. Hanna, Gerald S. Testing Students' Ability To Do Geometric Proofs: A Comparison of Three Objective Item Types. Journal for Research in Mathematics Education 2: 213-217; May 1971.

Multiple-choice items in which students selected either (1) what was given and what was proved or (2) the "reason"

were recommended over items which merely required the student to note whether a statement could be proved.

59. Hartlein, Marion L. Use of Items with Coded Numbers for Measuring Understanding of Elementary Mathematical Concepts. Arithmetic Teacher 13: 540-545; November 1966.

A multiple-choice test with matched items containing coded and non-coded numbers was designed to measure understanding of mathematical concepts. Coded items discriminated as well as non-coded items.

60. Henderson, George. Math Tests Analyzed. Wisconsin Journal of Education 100: 16-17, 27; May 1968.

Three mathematics tests were analyzed and items were classified for objectives tested.

61. Hendrickson, Gerry F. The Effect of Differential Option Weighting on Multiple-Choice Objective Tests. Journal of Educational Measurement 8: 291-296; Winter 1971.

The correlation of two mathematics subtests on the SAT decreased when Guttman weights were used to correct for guessing.

62. Horn, Dorothy M. Development of a pool of Mathematics Test Items for Grades 7 and 8. Arithmetic Teacher 16: 543-545; November 1969.

The development of an item pool is described, with several sample items.

63. Jeffery, Jay M. Psychological Set in Relation to the Construction of Mathematics Tests. Mathematics Teacher 62: 636-638; December 1969.

Students were found to develop a definite set toward problem solutions. Suggestions for classroom tests are discussed.

64. Koenker, Robert H. Measuring the Meanings of Arithmetic. Arithmetic Teacher 7: 93-96; February 1960.

Test items that evaluate understanding are illustrated and discussed.

65. Lankford, Francis G., Jr. What Can a Teacher Learn About a Pupil's Thinking Through Oral Interviews? Arithmetic Teacher 21: 26-32; January 1974.

Information gained from interviews in which seventh graders were asked to "say outloud your thinking as you compute" is presented.

66. Madden, Richard. New Directions in the Measurement of Mathematical Ability. Arithmetic Teacher 13: 375-379; May 1966.

Three criteria which should underlie the evaluation of mathematics achievement are discussed.

67. Martin, Wayne H. and Wilson, James W. The Status of National Assessment in Mathematics. Arithmetic Teacher 21: 49-53; January 1974.

The source, selection, and types of examples used in National Assessment, scoring procedures, and the nature of the sample are reported.

68. Merwin, Jack O. and Higgins, Martin J. Assessing the Progress of Education in Mathematics. Mathematics Teacher 61: 130-135; February 1968.

An overview of National Assessment in mathematics is given.

69. O'Brien, Thomas C. and Richard, June V. Interviews to Assess Number Knowledge. Arithmetic Teacher 18: 322-326; May 1971.

Interviews used to assess children's number knowledge are presented and discussed.

70. Payne, Joseph N. Giving the Student a Part in His Evaluation. Mathematics Teacher 50: 77-78; January 1957.

A student checklist for assessing progress in mathematics is presented.

71. Peck, Donald M. and Jencks, Stanley M. What the Tests Don't Tell. Arithmetic Teacher 21: 54-56; January 1974.

The need to have students demonstrate knowledge by using objects is discussed.

72. Pikaart, Len and Travers, Kenneth J. Teaching Elementary School Mathematics: A Simplified Model. Arithmetic Teacher 20: 332-342; May 1973.

A model for developing mathematics instruction in terms of goals, content, and teaching processes is presented.

73. Rappaport, David. Testing for Meanings in Arithmetic. Arithmetic Teacher 6: 140-143; April 1959.

Some practical suggestions for measuring understanding are presented.

74. Riedesel, C. Alan. Some Comments on Developing Proper Instrumentation for Research Studies in Mathematics. Arithmetic Teacher 15: 165-168; February 1968.

    Several considerations in developing, reading, and using studies in which research instruments are involved are suggested. Validity and effectiveness of such instruments are emphasized.

75. Romberg, Thomas A. and Braswell, James. Achievement Monitoring via Item Sampling: A Practical Data-Gathering Procedure for Formative Evaluation. Journal for Research in Mathematics Education 4: 262-270; November 1973.

    Details of item construction, periodic testing, and profile construction are described, as achievement monitoring and item sampling were used in evaluating the sixth-grade "Patterns in Arithmetic" program.

76. Romberg, Thomas A. and Wilson, James W. The Development of Mathematics Achievement Tests for the National Longitudinal Study of Mathematical Abilities. Mathematics Teacher 61: 489-495; May 1968.

    Development of the NLSMA tests is discussed in some detail.

77. Rusch, Reuben R.; Brown, John A.; and Delong, Arthur R. Meaning of an Arithmetic Test Score. Arithmetic Teacher 9: 145-148; March 1962.

    Analysis of test items and scores is discussed.

78. Sabers, Darrell L. and Feldt, Leonard. An Empirical Study of the Effect of the Correction for Chance Success on the Reliability and Validity of an Aptitude Test. Journal of Educational Measurement 5: 251-258; Fall 1968.

    Use of "guess" or "do not guess" directions had no significant effect on the predictive validity or reliability of an algebra aptitude test.

79. Schminke, Clarence W. The Arithmetic Folder. Arithmetic Teacher 9: 152-154; March 1962.

    The use of an arithmetic folder for systematic evaluation of behavior and work is described.

80. Swart, William L. Evaluation of Mathematics Instruction in the Elementary Classroom. Arithmetic Teacher 21: 7-11; January 1974.

    The need to evaluate the important objectives is discussed, with some specific illustrations.

81. Weaver, J. Fred. Big Dividends from Little Interviews. Arithmetic Teacher 2: 40-47; April 1955.

The use of interviews to evaluate mathematical growth is discussed. A sample interview record is given and implications for teaching suggested.

82. Wilkinson, Jack D. Teacher-directed Evaluation of Mathematics Laboratories. Arithmetic Teacher 21: 19-24; January 1974.

Specific suggestions are given for what and how to evaluate instruction in a mathematics laboratory setting, and how to use the results.

83. Wilson, James W. Standardized Tests Very Often Measure the Wrong Things. Mathematics Teacher 66: 295, 367-370; April 1973.

Limitations of standardized tests are discussed. The other side of the issue is discussed by Epstein (1973).

84. Wolff, Harry. Oral Testing. Mathematics Teacher 52: 384-387; May 1959.

The use of oral tests with algebra students is described.

IV. Dissertations: evaluation and tests in mathematics education

85. Bernabei, Raymond. A Logical Analysis of Selected Achievement Tests in Mathematics. (Western Reserve University, 1966.) Dissertation Abstracts 27A: 4121-4122; June 1967. [Order No. 67-4628]

A systematic approach to the analysis of standardized achievement tests using Bloom's Taxonomy and a comparison with goals of the SMSG program was presented.

86. Braswell, James Sidney, III. The Formative Evaluation of Patterns in Arithmetic, Grade 6, Using Item Sampling. (The University of Wisconsin, 1969.) Dissertation Abstracts International 31A: 672; August 1970. [Order No. 70-3484]

About 120 to 130 random responses to an item were sufficient to obtain a useful difficulty level for the selection of items.

87. Cotton, Timothy S. An Empirical Test of the Binomial Error Model Applied to Criterion-Referenced Tests. (University of Pittsburgh, 1971.) Dissertation Abstracts International 32A: 6186; May 1972. [Order No. 72-16,062]

The binomial model was found to be useful in developing tests for addition and subtraction.

88. Dahle, Mary McMahon. A Procedure for the Measurement of the Content Validity of Standardized Tests in Elementary Mathematics. (University of Southern California, 1970.) Dissertation Abstracts International 30A: 5336; June 1970. [Order No. 70-11,366]

    Using a grid of 120 objectives which ranged across five taxonomic levels, it was found that a textbook series corresponded closely to the distribution of objectives, while less correspondence was noted for standardized tests.

89. Donahue, Robert T. An Investigation of the Factor Pattern Involved in Arithmetic Problem Solving of Eighth Grade Girls. (The Catholic University of America, 1969.) Dissertation Abstracts International 30A: 2372; December 1969. [Order No. 69-19,720]

    Three factors were analyzed from a battery of tests on problem solving: a computational factor, a reading factor, and a reasoning factor.

90. Elsner, Priscilla Jo Edwards. A Study of Criterion-Referenced Assessment and Its Classroom Uses as Viewed by Teachers. (University of Northern Colorado, 1972.) Dissertation Abstracts International 33A: 3253-3254; January 1973. [Order No. 73-264]

    Teachers viewed the state assessment favorably and thought test results would be useful in planning instruction.

91. Evans, Robert Franklin. A Study of the Reliabilities of Four Arithmetic Attitude Scales and an Investigation of Component Mathematics Attitudes. (Case Western Reserve University, 1971.) Dissertation Abstracts International 32A: 3086; December 1971. [Order No. 72-32,182]

    The four scales had intercorrelations ranging from .59 to .83, indicating that a common construct was being sampled. Grade and concept effects were found to be significant.

92. Ferguson, Richard Leroy. The Development, Implementation, and Evaluation of a Computer-Assisted Branched Test for a Program of Individually Prescribed Instruction. (University of Pittsburgh, 1969.) Dissertation Abstracts International 30A: 3856; March 1970. [Order No. 70-4530]

    The described branched test could provide the same information for the mathematics unit studied as a conventional paper-and-pencil test in one-half the time and with substantially greater reliabilility in aiding instructional decision-making.

93. Futcher, Wilfred George. Scoring for Partial Knowledge in Mathematics Testing; A Study of a Modification and an Extension of Multiple-Choice Items Applied to the Testing of Achievement in Mathematics. (University of Toronto, 1969.) Dissertation Abstracts International 31A: 1619-1620; October 1970. [Order from National Library of Canada at Ottawa.]

    Partial-knowledge scoring methods were found to be more discriminating and reliable than conventional right-wrong methods. Extended format in which multiple choice was offered at several stages was recommended.

94. Graham, Glenn Thomas. Sequentially Scaled Mathematics Achievement Tests: Construction Methodology and Evaluation Procedures. (University of Pittsburgh, 1966.) Dissertation Abstracts 27A: 3308; April 1967. [Order No. 67-4567]

    Use of scalogram analysis was made in constructing tests in five areas of arithmetic achievement.

95. Gridley, John David, Jr. An Empirical Investigation of the Construct of Mathematics Achievement in the Elementary Grades Based on the Method of Homogeneous Keying. (Fordham University, 1971.) Dissertation Abstracts International 32A: 1914; October 1971. [Order No. 71-27,010]

    Mathematics achievement as measured by a standardized test was found to consist of several empirically defined clusters of items, which varied from grade to grade. The meaningfulness of the total score was questioned.

96. Gupta, Ram Krishna. Interactions of Achievement Test Items in the Context of Repeated Measurements of Groups, Using Different Mathematics Texts. (University of Minnesota, 1967.) Dissertation Abstracts 28A: 2093-2094; December 1967. [Order No. 67-14,613]

    Analyses of student responses to groups of items indicated differences or interactions between items, teachers, texts, and sexes.

97. Impara, James Clement. An Experimental Comparison of Matrix Sampling and Examinee Sampling for Estimating Test Norms for Different Target Groups on Different Types of Tests. (The Florida State University, 1972.) Dissertation Abstracts International 33A: 4942; March 1973. [Order No. 73-4689]

    Matrix sampling appeared feasible for estimating mean achievement in the two test domains studied with both disadvantaged and non-disadvantaged pupils in grade 4.

98. Kwansa, Kofi Bassa. Investigation of the Relative Content Validity of Norm-Referenced and Domain-Referenced Arithmetic Tests. (University of Pittsburgh, 1972.) Dissertation Abstracts International 33A: 3959-3960; February 1973. [Order No. 73-4153]

    The domain-referenced tests had higher content validity than did the norm-referenced tests. Scores on the two forms correlated highly.

99. Macready, George Byron. An Investigation into the Nature of Interitem Relations and the Structure of Domain Hierarchies Found Within a Domain Referenced Testing System. (University of Minnesota, 1972.) Dissertation Abstracts International 33A: 2174; November 1972. [Order No. 72-27,776]

    Little variability in items was found within the various domains of items. Two item-generation procedures produced quite similar hierarchies. It appeared feasible to test students on a sample of items and infer how they would perform on the domain.

100. Martin, Mavis Doughty. Reading Comprehension, Abstract Verbal Reasoning, and Computation as Factors in Arithmetic Problem Solving. (State University of Iowa, 1963.) Dissertation Abstracts 24: 4547-4548; May 1964. [Order No. 64-3395]

    High correlations among reading, ability, and computation scores were found, indicating a complex interaction and the cruciality of all to problem-solving skill.

101. Mastantuono, Albert Kenneth. An Examination of Four Arithmetic Attitude Scales. (Case Western Reserve University, 1970.) Dissertation Abstracts International 32A: 248; July 1971. [Order No. 71-19,029]

    The correlations of four attitude scales with grade and sex were analyzed. Scores on two scales correlated significantly with achievement.

102. Montague, Margariete Ann. Use of Matrix Sampling Procedures with Selected Examinee and Item Populations to Assess Achievement in Mathematics. (The University of Wisconsin, 1971.) Dissertation Abstracts International 32A: 5475; April 1972. [Order No. 72-2651]

    The feasibility of concurrently and randomly sampling examinees and items to obtain group data generalizable to a universe of each was established.

103. Pruzek, Robert Marshall. A Comparison of Two Methods for Studying Test Items. (The University of Wisconsin, 1967.) Dissertation Abstracts 28A: 3035; February 1968. [Order No. 67-12,464]

A method of categorizing test items was used with a mathematics standardized test (CEEB).

104. Purcell, Joseph E. The Relation of Student Classroom Marks and Regents Examination Marks to Teacher Knowledge of Student Standardized Test Scores. (State University of New York at Albany, 1972.) Dissertation Abstracts International 33A: 2722; December 1972. [Order No. 72-31,804]

No significant difference in student marks were found when mathematics, English, and history teachers were given standardized test data or only urged to seek it.

105. Pyrczak, Fred, Jr. Objective Evaluation of the Quality of Multiple-Choice Test Items. (University of Pennsylvania, 1972.) Dissertation Abstracts International 33A: 3401; January 1973. [Order No. 73-1436]

The conventional discrimination index appeared to be a moderately valid measure of item quality, though a substantial amount of interrater variability remained non-explained.

106. Riley, Sister Mary Felicitas. An Analysis of Timed and Untimed Test Scores of Subjects from Two Different Arithmetic Curricula. (Fordham University, 1963.) Dissertation Abstracts 25: 2354-2355; October 1964. [Order No. 64-2432]

Students tended to do significantly better on the untimed test when it was administered after the timed test.

107. Romberg, Thomas Albert. Derivation of Subtests Measuring Distinct Mental Processes Within the NLSMA Algebra Achievement Test. (Stanford University, 1968.) Dissertation Abstracts 29A: 419; August 1968. [Order No. 68-11,341]

Three scaling methods were used to derive subtests of general mental processes associated with mathematics.

108. Spilman, Helen W. The Use of a Single Item-Sample to Estimate Group Achievement. (The City University of New York, 1973.) Dissertation Abstracts International 32A: 177-178; July 1973. [Order No. 73-14,385]

None of the estimated means derived from mini-tests of items randomly drawn from a standardized achievement test were within one standard error of measurement of the total test means.

109. Walker, Charles Everett. The Effect of Variations in Test Administration Conditions on Arithmetic Test Performance. (The University of Rochester, 1969.) Dissertation Abstracts International 31A: 242-243; July 1970. [Order No. 70-2946]

No differences in scores on a computation test were found that could be attributed to variations in the use of separate answer sheets, scrap paper, or test booklets.

V. ERIC documents: evaluation and tests in mathematics education

110. Alkin, Marvin C. and others. Mathematics K-3, Instructional Objectives Exchange. Los Angeles: Center for the Study of Evaluation, 1969. ERIC: ED 035 568. 190 pages.

111. Alkin, Marvin C. and others. Mathematics 4-6, Instructional Objectives Exchange. Los Angeles: Center for the Study of Evaluation, 1969. ERIC: ED 034 702. 250 pages.

112. Alkin, Marvin C. and others. Mathematics 7-9, Instructional Objectives Exchange. Los Angeles: Center for the Study of Evaluation, 1969. ERIC: ED 035 567. 282 pages.

Each of the three above documents present objectives and items for the specified level.

113. Ash, Michael J. and Sattler, Howard E. A Video Tape Technique for Assessing Behavioral Correlates of Academic Performance. March 1973. ERIC: ED 074 747. 18 pages.

The relationship between videotape-based observer judgments of attention to task and paper-and-pencil measures of academic performance was studied; data supported the use of indirect observational methods in assessing school performance.

114. Ascher, Gordon. Individualized Instruction and Statewide Assessment: The New Jersey Educational Assessment Program. February 1973. ERIC: ED 074 129. 22 pages.

Goals, test development and administration, and test results and their use are discussed.

115. Besel, Ronald. Using Group Performance to Interpret Individual Responses to Criterion-Referenced Tests. February 1973. ERIC: ED 076 658. 10 pages.

The contention that interpretation of a student's performance on a criterion-referenced test should be independent of the performance of his classmates is challenged, and an alternative model proposed.

116. Cahen, Leonard S. and others. A Comparison of School Mean Achievement Scores with Two Estimates of the Same Scores Obtained by the Item-Sampling Technique. November 1970. ERIC: ED 052 241. 26 pages.

Reasonably close estimates of mean performance were obtained from the item-sampling situation as compared to means estimated from the conventional type of testing.

117. Callahan, Leroy G. Clinical Evaluation and the Classroom Teacher. 1973. ERIC: ED 076 640. 8 pages.

Some of the potential of clinical evaluation procedures in making judgments on student learning in elementary school mathematics is examined. Use of a clinical interview with videotaping procedures is emphasized.

118. Donovan, David and others. Objectives and Procedures: The First Report of the 1972-73 Michigan Educational Assessment Program. Lansing: Michigan State Dept. of Education, October 1972. ERIC: ED 073 139. 36 pages.

Objectives of the assessment and procedures to be used are described.

119. Chandler, Arnold M. and others. Guidelines to Mathematics, K-6. Key Content Objectives, Student Behavioral Objectives, and Other Topics Related to Elementary School Mathematics. Madison: Wisconsin State Dept. of Public Instruction, 19__. ERIC: ED 051 185. 58 pages.

Course content and related behavioral objectives are presented, with suggestions for developing tests and choosing appropriate methods of evaluation.

120. Coolsby, Thomas M., Jr. Evaluation of Cognitive Development: An Observational Technique -- Pre-Mathematics Skills Inventory. Athens, Georgia: Research and Development Center in Educational Stimulation, June 1969. ERIC: ED 046 989. 33 pages. Available from EDRS only in microfiche.

The development of an instrument for evaluating cognitive growth in preprimary children by means of an observational inventory is described.

121. Harris, Margaret L. and Harris, Chester W. Newly Constructed Reference Tests for Cognitive Abilities. Working Paper No. 80. Madison: Wisconsin Research and Development Center for Cognitive Learning, November 1971. ERIC: ED 072 114. 144 pages.

This document presents 35 tests that were constructed or adapted for inclusion in a battery of reference tests for cognitive abilities appropriate for the fourth- and fifth-grade level.

122. Harris, Margaret L. and Romberg, Thomas A. Measuring Mathematics Concept Attainment: Boys and Girls. Technical Report 195. Madison: Wisconsin Research and Development Center for Cognitive Learning, November 1971. ERIC: ED 070 659. 34 pages.

Test development efforts for constructing 12 tests to measure achievement of 30 selected mathematics concepts of sets, division, and expressing relationships are described. Item statistics are discussed.

123. Harris, Margaret L. and Romberg, Thomas A. An Analysis of Content and Task Dimensions of Mathematics Items Designed to Measure Level of Concept Attainment. Technical Report 196. Madison: Wisconsin Research and Development Center for Cognitive Learning, November 1971. ERIC: ED 070 660. 42 pages.

Analysis of the items discussed in ED 070 659, administered to fifth- and sixth-graders, is discussed.

124. Harsh, J. Richard. Diagnostic Mathematics (Form A, Form B, and Test Manual). Fort Worth, Texas: Fort Worth Independent School District and National Consortia for Bilingual Education, 1972. ERIC: ED 062 174. 36 pages.

The test provides a measure of the conventional sequence of arithmetic computation and selected applications. Each form consists of 44 completion items, with space for computation.

125. Harvill, Leo M. Evaluation of Several Methods for Measuring Young Children's Educational Attitudes. Vermillion: South Dakota University, May 1971. ERIC: ED 056 059. 38 pages.

Five scales designed to measure the attitudes of young children toward arithmetic, reading, and art are described.

126. Henderson, George L. and others. Guidelines to Mathematics, 6-8. Key Content Objectives, Student Behavioral Objectives, and Other Topics Related to Grade 6-8 Mathematics. Madison: Wisconsin State Dept. of Public Instruction, 19__. ERIC: ED 051 186. 44 pages.

Behavioral objectives for 17 mathematical concepts are included.

127. Henderson, George L. and others. Wisconsin Statewide Assessment Mathematics: An Exemplary Mathematics Program, Grades K-8, and a Hierarchy of Student Behavioral Objectives K-8. Madison: Wisconsin State Dept. of Education, 19__. ERIC: ED 069 475. 38 pages.

Overall goals and a hierarchy of over 400 mathematics content objectives are listed in a prerequisite and sequential order and also organized in a grid form, with suggestions for use.

128. Kissel, Mary Ann and Yeager, John L. An Investigation of the Efficiency of Various Observational Procedures. February 1971. ERIC: ED 048 372. 28 pages.

Factors related to length and time of observations were studied to determine their relative efficiency.

129. Knipe, Walter H. and Krahmer, Edward F. An Application of Criterion Referenced Testing. February 1973. ERIC: ED 074 154.

Mathematics criterion-referenced tests for grades 3 through 9 are described.

130. Kriewell, Thomas E. and Hirsch, Edward. The Development and Interpretation of Criterion-Referenced Tests. Madison: Wisconsin Research and Development Center for Cognitive Learning, February 1969. ERIC: ED 042 815. 26 pages.

The use of a strict item-sampling model for constructing criterion-referenced tests is discussed.

131. Lieberman, Marcus and others. Primary Mathematics: Behavioral Objectives and Test Items. Downers Grove, Illinois: Institute for Educational Research, 1972. ERIC: ED 066 494. 173 pages.

132. Lieberman, Marcus and others. Intermediate Mathematics: Behavioral Objectives and Test Items. Downers Grove, Illinois: Institute for Educational Research, 1972. ERIC: ED 066 495. 587 pages.

133. Lieberman, Marcus and others. Junior High Mathematics: Behavioral Objectives and Test Items. Downers Grove, Illinois: Institute for Educational Research, 1972. ERIC: ED 066 496. 236 pages.

134. Lieberman, Marcus and others. High School Mathematics: Behavioral Objectives and Test Items. Downers Grove, Illinois: Institute for Educational Research, 1972. ERIC: ED 066 497. 810 pages.

Each of the four above documents present the Objective-Item Bank for the specified level.

135. McNaughton, A. E. Evaluation in Secondary School Mathematics. Victoria, Australia: Victoria Education Department, 1970. ERIC: ED 065 275. 22 pages.

The development of 20 objective test items is traced.

136. Norris, Eleanor L. and Bowes, John E. (Eds.). National Assessment of Educational Progress, Mathematics Objectives. Ann Arbor, Michigan: National Assessment of Educational Progress, 1970. ERIC: ED 063 140. 41 pages.

No illustrative test items are included, but the general nature of three levels of tasks is described and the specific topics for each age to be tested are listed.

137. Patalino, Marianne. Rationale and Use of Content-Relevant Achievement Tests for the Evaluation of Instructional Programs. Los Angeles: Center for the Study of Evaluation, May 1970. ERIC: ED 041 044. 52 pages.

Problems in current course evaluation methods are discussed and an alternative method is described for the construction, analysis, and interpretation of a test to evaluate instructional programs. Two forms of the Diagnostic Test for the Los Angeles Model Mathematics Project are included.

138. Poggio, John P. and Glasnapp, Douglas R. Item-Sampling as a Classroom Evaluation Technique. 1973. ERIC: ED 076 692. 9 pages.

Item-sampling can be employed for classroom assessment, providing feedback over a greater range of content objectives than can be tested by typical test construction.

139. Romberg, Thomas A.; Shepler, Jack L.; and Wilson, James W. Three Experiments Involving Probability Measurement Procedures with Mathematics Test Items. Technical Report No. 129. Madison: Wisconsin Research and Development Center for Cognitive Learning, 1970. ERIC: ED 044 315. 33 pages.

For multiple-choice tests (with items from the NLSMA bank of "insightful" items), reliability coefficients did not increase when the student was asked to specify a belief in the probability of each of the given alternatives being correct.

140. Romberg, Thomas A. and Steitz, Jean. Items to Test Level of Attainment of Mathematics Concepts by Intermediate Grade Children. Working Paper No. 56. Madison: Wisconsin Research and Development Center for Cognitive Learning, November 1971. ERIC: ED 070 653. 76 pages.

A twelve-part paradigm for testing level of concept attainment was used to construct 353 items covering the topics of sets, division, and expressing relationships.

141. Rothney, John W. M. Evaluating and Reporting Pupil Progress. What Research Says to the Teacher Series, No. 7. Washington: Association of Classroom Teachers, National Education Association, 1972. ERIC: ED 079 240. 36 pages. Available from EDRS only in microfiche.

Standard criteria for assessing achievement, procedures to assess personal-social development, and the time for evaluation are included.

142. Roudabush, Glenn E. and Green, Donald Ross. Some Reliability Problems in a Criterion-Referenced Test. February 1971. ERIC: ED 050 144. 13 pages.

The development of a 400-item Prescriptive Mathematics Inventory is discussed.

143. Schmeiser, Cynthia Board and Whitney, Douglas R. The Effect of Selected Poor Item Writing Practices on Test Difficulty, Reliability and Validity: A Replication. 1973. ERIC: ED 075 498. 15 pages.

The effect of violating four selected principles of writing multiple-choice items is described.

144. Stewart, Deborah Miller. Development of a Group Test of Arithmetic Achievement for Developing Mathematical Processes, Arithmetic Book 1. Madison: Wisconsin Research and Development Center for Cognitive Learning, August 1970. ERIC: ED 047 979. 147 pages.

The construction of a group-administered test of the knowledge of kindergarten and first-grade children for the content of the "Developing Mathematical Processes" program is described.

145. Suydam, Marilyn N. Unpublished Instruments for Evaluation in Mathematics Education: An Annotated Listing. Columbus, Ohio: ERIC Information Analysis Center for Science, Mathematics, and Environmental Education, January 1974. ERIC: SE 017 118. 259 pages.

Non-commercial investigator-developed tests and other instruments to assess mathematical instruction, reported in journals and dissertations from 1964 through 1973, are listed. For approximately 200 instruments, information on content, format, sample, reliability, correlations, and validity is included, as well as references. Other instruments for which only partial information was available are also cited. (No instruments are included.)

146. Williams, S. Irene and Jones, Chancey O. A Comparison of Interview and Normative Analysis of Mathematics Questions. Princeton, New Jersey: Educational Testing Service, April 1972. ERIC: ED 067 397. 49 pages.

Interview questions can provide useful information to supplement other forms of evaluation.

147. Womer, Frank B. What is National Assessment? Ann Arbor, Michigan: National Assessment of Educational Progress, 1970. ERIC: ED 067 394. 56 pages.

This is a general description of the plan for National Assessment, a systematic, census-like survey of knowledges, skills, understandings, and attitudes designed to sample four age levels in ten different subject areas.

148. ---. Annual Mathematics Examination, 1966-1972. Lincoln, Nebraska: Committee on High School Contests, Mathematical Association of America, 1972. ERIC: ED 070 634. 97 pages. Available from EDRS only in microfiche.

Sets of the annual examination are included, with solution keys.

149. ---. Sixth Grade Mathematics: A Needs Assessment Report. Austin: Texas Education Agency, 1972. ERIC: ED 071 879. 132 pages.

The basic objectives are summarized, and the percentage of sixth graders mastering each objective is given.

150. ---. Staff Utilization for Continuous Progress Education: Math Pretests and Posttests for Third and Fourth Grades. Phoenix, Arizona: Scottsdale Public Schools, 1973. ERIC: ED 077 771. 444 pages.

This is a collection of mathematics pre- and posttests for grades 3 and 4 on sets, place value, addition-subtraction, multiplication, division, multiplication-division, and fractions.

- - - - - - - - - - -

Note: The references cited are not available from ERIC/SMEAC. "Order No." in references for dissertations pertains to the number to be used in ordering a copy of the dissertation from University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106. (Costs are $4 for microfilm and $10 for Xerography copy.) ERIC documents may be ordered from the Educational Document Reproduction Service, P. O. Drawer O, Bethesda, Maryland 20014. Specify the ED number. (Costs are $0.65 for microfiche and $3.29 per hundred pages or any part thereof for hard copy.)

ERIC Document Reproduction Service

References in this paper which carry ED six-digit numbers are available through the ERIC Document Reproduction Service (EDRS). Duplicates of papers are produced in two formats. Microfiche duplicates are reduced photocopies which require the use of a microfiche reader. Such readers are now widely available through most libraries. Hard-copy duplicates are full-sized duplicates reproduced through a Xerox process.

The current price schedule for duplicates can always be found in the most recent issue of Research in Education. The price schedule in effect at the time of this printing is:

MICROFICHE DUPLICATE:

On demand, by title	$ 0.65

HARD COPY:

On demand, by title,

1 - 100 pages	$ 3.29
101 - 200 pages	$ 6.58
201 - 300 pages	$ 9.87
301 - 400 pages	$13.16
401 - 500 pages	$16.45
each additional 1 - 100 page increment	$ 3.29

Orders should be sent to:

ERIC Document Reproduction Service
P. O. Drawer O
Bethesda, Maryland 20014